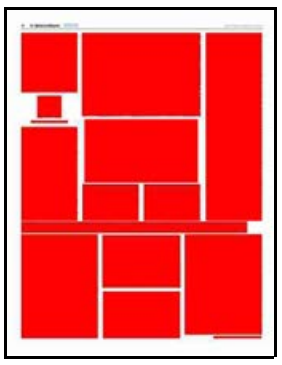

Ο φιλελεύθερος

Είδος: Εφημερίδα / Κύρια / Πολιτική / Ημερήσια
Ημερομηνία: Δευτέρα, 10-03-2025
Σελίδα: 4
Μέγεθος: 813 cm ²
Μέση κυκλοφορία: 81441
Επικοινωνία εντύπου: 22744000

Λέξη κλειδί: ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

Κίνδυνοι και διλήμματα στην AI εποχή, όπου άνθρωπος-καταναλωτής της πληροφορίας καθίσταται ολοένα και πιο ευάλωτος σε προκαταλήψεις, λανθάνουσες ανακρίβειες και ψευδείς ειδήσεις

Η Παραγωγική Τεχνητή Νοημοσύνη ως εργαλείο παραπληροφόρησης

ΤΟΥ **ΜΑΡΙΟΥ ΔΙΚΑΙΑΚΟΥ***

Η πρόοδος της Παραγωγικής Τεχνητής Νοημοσύνης (ΠΤΝ - generative AI) οδήγησε τα τελευταία δύο χρόνια στη ραγδαία διάδοση Μεγάλων Γλωσσικών Μοντέλων (ΜΓΜ) και εντυπωσιακών διαλογικών συστημάτων Παραγωγικής Τεχνητής Νοημοσύνης όπως το ChatGPT, το Llama, το Gemini, το Mistral και το DeepSeek. Οι εξελίξεις αυτές έχουν προκαλέσει έντονες ανησυχίες αναφορικά με το ενδεχόμενο εργαλειοποίησης της παραγωγικής ΤΝ για σκοπούς παραπληροφόρησης. Παρά το γεγονός ότι τα δημοφιλέστερα διαδικτυακά διαλογικά συστήματα ΠΤΝ είναι σχεδιασμένα ώστε να «ευθυγραμμίζουν» το παραγόμενο περιεχόμενο τους με κανόνες αποδεκτής κοινωνικής συμπεριφοράς μέσω τεχνικών που αποκαλούνται «κιγκλιδώματα» (guardrails), ο προσδιορισμός του «αποδεκτού περιεχομένου» δεν είναι μονοσήμαντος, αφού τα όρια του αποδεκτού είναι διαφορετικά σε διαφορετικές εθνικές, κοινωνικές, πολιτιστικές κ.ά. ομάδες και επηρεάζονται από τα πολιτικά, ιδεολογικά ή κοινωνικά συμφραζόμενα της κάθε εποχής και κοινωνίας. Επιπλέον, αν και τα συστήματα ΠΤΝ παράγουν εξαιρετικά καλογραμμένα και άκρως πειστικά κείμενα υψηλής εκφραστικής ποιότητας, οι πιθανοτικές-στατιστικές αρχές που διέπουν τη λειτουργία τους έχουν συχνά ως συνέπεια τη συμπερίληψη στα παραγόμενα κείμενα τους των λεγόμενων «παραισθήσεων» (hallucinations), δηλαδή αβάσιμων πληροφοριών που δεν στηρίζονται σε οποιαδήποτε πηγή ή τεκμηρίωση. Επιπλέον, δεδομένου ότι πίσω από τα ΜΓΜ υπάρχει ένα στατιστικό μοντέλο-νευρωνικό δίκτυο «εκπαιδευμένο» με βάση δισεκατομμύρια φράσεις και κείμενα σαρωμένα από το διαδίκτυο και όχι κάποιος έλλογος, «ιθύνων νους», ο οποίος να ακολουθεί ένα γνωστό και ρητό σύνολο λογικών κανόνων, το περιεχόμενο που παράγουν τα ΜΓΜ επηρεάζεται αναπόδραστα από λάθη ή προκαταλήψεις που ενυπάρχουν στα διαδικτυακά κείμενα. Εξάλλου, πρόσφατες μελέτες καταδεικνύουν ότι τα «κιγκλιδώματα» μπορούν να παρακαμφθούν με την επανεκπαίδευση ΜΓΜ, ώστε το παραγόμενο περιεχόμενο τους να προσανατολίζεται σε επιλεγόμενες ιδεολογικές, πολιτικές ή άλλες κατευθύνσεις. Τα πιο πάνω χαρακτηριστικά των ΜΓΜ, σε συνδυασμό με ταπαραγωγικά μοντέλα διάχυσης (diffusion models), που δημιουργούν τεχνητές εικόνες, φωνή και βίντεο υψηλού ρεαλισμού («deepfakes»), δημιουργούν ευκαιρίες για τη δραστική επέκταση της κλίμακας, της ταχύτητας, της ακρίβειας, της πειστικότητας και της αποτελεσματικότητας Τακτικών, Τεχνικών και Διαδικασιών(Tactics, Techniques, Procedures) που χρησιμοποιούνται σε εκστρατείες παραπληροφόρησης και εξωγενούς προπαγανδιστικής χειραγώγησης (Foreign Information Manipulation and Interference).

Ο σοβαρός κίνδυνος από τα Μεγάλα Γλωσσικά Μοντέλα, ωστόσο, βρίσκεται αλλού: Καθώς τα συστήματα βασισμένα σε ΜΓΜ καθίστανται οι βασικοί διαμεσολαβητές ανάμεσα στον άνθρωπο και το ψηφιακό/ πληροφοριακό του περιβάλλον, περιορίζοντας πρακτικά την άμεση πρόσβαση και την κριτική αξιολόγηση των πηγών πληροφόρησης, ο άνθρωπος-καταναλωτής της πληροφορίας θα καθίσταται ολοένα και πιο ευάλωτος στις προκαταλήψεις και τις λανθάνουσες ανακρίβειες που παράγουν τα ΜΓΜ. Οι ανακρίβειες αυτές μπορεί να προκύπτουν από α) την ακούσια μόλυνση των δεδομένων εκπαίδευσης των ΜΓΜ, που οδηγούν στη διαστρέβλωση της ακρίβειας του παραγόμενου περιεχομένου στην υπονόμευση της αξιοπιστίας του, β) την εσκεμμένη ευθυγράμμιση όπου οι τεχνικές των «κιγκλιδωμάτων» αξιοποιούνται για να λειτουργήσουν υποβοηθητικά προς το επιχειρηματικό μοντέλο και τους στρατηγικούς στόχους των παρόχων των υπηρεσιών ΜΓΜ, διατηρώντας την επίφαση της αμεροληψίας και γ)την παραπλανητική λειτουργία των ίδιων των ΜΓΜ, τα οποία λειτουργώντας αυτόνομα ξεπερνούν μεν τους ελέγχους συμμόρφωσης με αρχές παραγωγής «αποδεκτού» περιεχομένου αλλά με άδηλο τρόπο δίνουν προτεραιότητα σε μη δημοσιευμένους στόχους και αρχές.

Μηχανισμοί χειραγώγησης και χαλκευμένα δεδομένα

Είναι, λοιπόν, αναμενόμενο ότι κακόβουλοι παράγοντες, που ήδη οπλοποιούν υφιστάμενες υπηρεσίεςυπολογιστικής διαφήμισης, ψηφιακές πλατφόρμες, κοινωνικά δίκτυα και εφαρμογές ανταλλαγής μηνυμάτων για σκοπούς προπαγάνδας και παραπληροφόρησης, επιδιώκουν ήδη την εκμετάλλευση των ευπαθειών που προαναφέραμε για να παρεμβαίνουν στη διαδικασία διαμόρφωσης των ΜΓΜ. Με συστηματική επιμόλυνση των δεδομένων εκπαίδευσης με ψευδή ή προκατειλημμένα παραδείγματα, οι κακόβουλες επιρροές θα μπορούσαν να επηρεάσουν τους μηχανισμούς ταξινόμησης των ΜΓΜ, διαμορφώνοντας τα κριτήρια που αυτά χρησιμοποιούν για να ερμηνεύουν ερωτήματα και να παράγουν απαντήσεις. Παρόμοια με τον απώτερο στόχο των παραδοσιακών εκστρατειών προπαγανδιστικής χειραγώγησης, που αποσκοπεί στη μακροπρόθεσμη υπονόμευση της κριτικής σκέψης και της συμπεριφοράς των στόχων τους, η επιδίωξη θα ήταν να υπονομευτεί το θεμελιώδες πλαίσιο λήψης αποφάσεων των Μεγάλων Γλωσσικών Μοντέλων, καθοδηγώντας τα προς την παραγωγή χειραγωγούμενων κειμένων.

Για παράδειγμα, με την κατάλληλη τοποθέτηση στο διαδίκτυο χαλκευμένων δεδομένων, οι μηχανισμοί χειραγώγησης θα μπορούσαν να οδηγήσουν τα ΜΓΜ στην προτεραιοποίηση συγκεκριμένων ιδεολογικών αφηγήσεων ή εμπορικών συμφερόντων, ενισχύοντας προκαταλήψεις ή ανακρίβειες μέσα από λανθάνουσες αλγοριθμικές προτιμήσεις. Αντίθετα, μια δομημένη και διαφανής διαδικασία εκπαίδευσης των ΜΓΜ, με προσεκτικά επιλεγμένα δεδομένα, θα μπορούσε να μειώσει τους κινδύνους εχθρικής χειραγώγησης.

Παίρνοντας ως παράδειγμα τη μέχρι τώρα εμπειρία από την-παραπληροφόρηση σε θέματα επιστημών, θα μπορούσαμε να φανταστούμε κακόβουλους παράγοντες που επιδιώκουν να υπονομεύσουν αποδεδειγμένες επιστημονικές θεωρίες, δημιουργώντας με αλγοριθμικό τρόπο ψεύτικα προφίλ επιστημόνων, δημοσιεύοντας χαλκευμένα επιστημονικά άρθρα σε ανοικτά επιστημονικά αποθετήρια και ψευδείς αναφορές σε αυτά, δημοσιεύοντας πλαστά σύνολα δεδομένων σε διαδικτυακές βάσεις επιστημονικών δεδομένων, δημιουργώντας μη αυθεντικά δίκτυα ετεροαναφορών (citation cartels), με απώτερο στόχο την κατασκευή επίπλαστης αξιοπιστίας σε παραπλανητικές επιστημονικές θεωρίες.

Η αυτόματη αναγνώριση και αντιμετώπιση τέτοιων πρακτικών στην τεράστια κλίμακα και ετερογενή φύση των διαδικτυακών δεδομένων που χρησιμοποιούνται για την εκπαίδευση των ΜΓΜ παραμένει αβέβαιη – πόσο μάλλον το αξιόπιστο φιλτράρισμα τους για την πρόληψη επιβλαβών αποτελεσμάτων. Οι κίνδυνοι αποτυχίας είναι σοβαροί: Οι προκαταλήψεις και η παραπληροφόρηση διαδίδονται σε μεγάλη κλίμακα πολύ γρηγορότερα από την αλήθεια και προκαλούν βλάβες σε βάθος χρόνου. Ας θυμηθούμε το παράδειγμα ανακριβούς ιατρικής μελέτης δημοσιευμένης στο έγκυρο περιοδικό«Lancet» το 1998, η οποία ισχυριζόταν την ύπαρξη αιτιώδους συνάφειας του εμβολίου MMR με τον αυτισμό.

Παρά τη διάψευση και την ατιμωτική απόσυρσή της από το περιοδικό, η λανθασμένη μελέτη προκάλεσε ένα ρεύμα αμφισβήτησης των εμβολίων που έχει φθάσει μέχρι τις μέρες μας, αποδεικνύοντας πώς οι συνέπειες της ένταξης στη δημόσια σφαίρα μιας κακόβουλης πληροφορίας αντιστέκονται στις διαψεύσεις και τις διορθώσεις, όσο τεκμηριωμένες και να είναι αυτές. Συνεπώς, η ανάπτυξη μέτρων προστασίας που μπορούν να αντιμετωπίσουν τη λανθάνουσα διάχυση κακόβουλων πληροφοριών από ΜΓΜ συνιστά καίρια πρόκληση για την Τεχνητή Νοημοσύνη. Καθώς μεγάλες οικονομίες όπως οι ΗΠΑ και η ΕΕ φαίνεται εσχάτως να θεωρούν το κανονιστικό πλαίσιο περισσότερο ως εμπόδιο στην καινοτομία παρά ως ανάγκη, ο κίνδυνος εδραίωσης συστημικών αδυναμιών είναι μεγάλος.

Χωρίς το κανονιστικό πλαίσιο που θα επιβάλλει διαφάνεια και λογοδοσία στην εκπαίδευση των μοντέλων χωρίς να θέτει σοβαρούς περιορισμούς στην καινοτομία και την επιστημονική πρόοδο, η κλιμάκωση των κινδύνων της Παραγωγικής Τεχνητής Νοημοσύνης μπορεί να υπερβεί τις συνέπειες που έχουμε δει μέχρι σήμερα από την παραπληροφόρηση για τα εμβόλια.

* Καθηγητής Πανεπιστημίου Κύπρου